Text Mining for Historical Documents
Metadata, Standardisation, Semantic Web

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Wintersemester 2010/11

22.02.2011

Metadata

Data and Metadata

Distinction

primary data (=data) vs. secondary data (=metadata)

Types of metadata

- data that contextualises the object (e.g., when and where found; written by whom etc.)
  - created manually by experts (expedition leaders, curators etc.)
  - usually created immediately
  - may or may not be digitised
  - typically fixed / sacrosanct, can only be (monotonically) added to (at least for museums)
    $\Rightarrow$ **metadata provenance**
- data that provides additional information to improve data access (semantic annotation in the widest sense)
  - manually, semi-automatically, or automatically created
  - typically digitised
  - can change over time (i.e., monotonically increase)

$\Rightarrow$ As a rule of thumb, human generated/verified metadata should not be deleted or overwritten.

‘Semantic’ Metadata (1)

Types of metadata annotations

'Semantic' Metadata (1)

Types of metadata annotations

- general linguistic analyses, esp.
  - word sense disambiguation
  - co-reference resolution
  - named entity tagging
  - named entity disambiguation and linking
- enrichment for information retrieval
  - synonyms (possibly hypernyms, hyponyms)
  - annotation with modern language equivalents (for words from older language varieties)
  - annotation with corrected forms (for OCR errors or typos)
  - translations into other languages
  - transcripts of speech or non-OCRed material
  - content annotation (keywords, descriptions)

'Semantic' Metadata (2)

Types of metadata annotations (contd)

- data provenance information, e.g.:
  - was an entry in a database corrected (when? by whom? how?)
  - was additional information entered (when? by whom? how?)
- miscellanea, e.g.:
  - information about links between data sources
  - information extraction information (e.g., explicit structuring of semi-structured data)
  - browsing history (e.g., for website content)

$\Rightarrow$ metadata can come in several layers

Standardisation

Controlled Vocabularies

Motivation

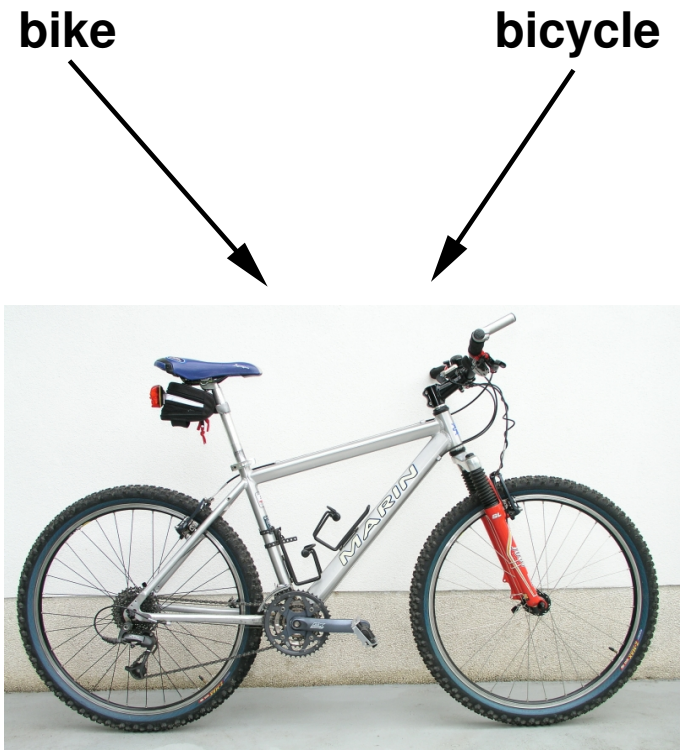
natural language can be fuzzy (**synonymy** and **polysemy**)

$\Rightarrow$ many-to-many mapping between form and meaning

Motivation

natural language can be fuzzy (**synonymy** and **polysemy**)

$\Rightarrow$ many-to-many mapping between form and meaning

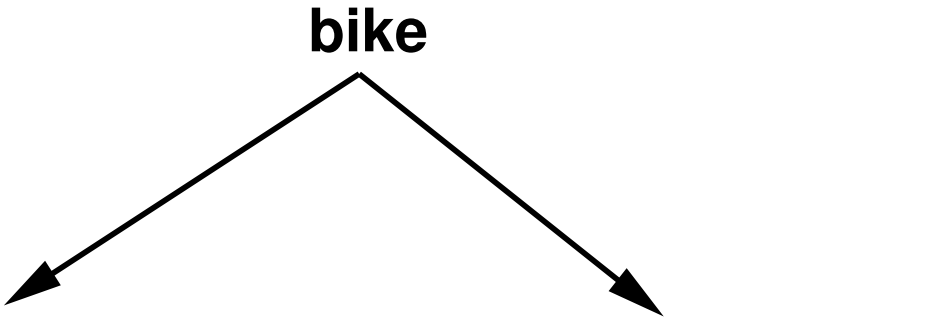


(Source: http://en.wikipedia.org/wiki/File:Marin_bike.jpg)

Motivation

natural language can be fuzzy (**synonymy** and **polysemy**)

$\Rightarrow$ many-to-many mapping between form and meaning



(Source: http://en.wikipedia.org/wiki/File:Marin_bike.jpg)

(Source: http://en.wikipedia.org/wiki/File:Triumph_T_110_650_cc_1954.jpg)

Controlled Vocabularies

Motivation

natural language can be fuzzy (**synonymy** and **polysemy**)
$\Rightarrow$ many-to-many mapping between form and meaning

Controlled Vocabularies (CVs)

- in keyword-based search synonymy and polysemy lower recall and precision, respectively
- controlled vocabularies fix which terms can be used for annotation and searching
  $\Rightarrow$ avoid ambiguity and impreciseness
- CH institutes use existing CVs (e.g., domain thesauri) or develop CV inhouse
- may or may not improve retrieval results, depending on situation (Svenonius, 1986)

Semantic Interoperability

What?

In order to (automatically) share information across collections and/or institutes, the semantic metadata have to be compatible. Controlled vocabularies (=lists of standardised terms) are not sufficient.

Solution

metadata standards in the form of ontologies or domain descriptions, e.g.:

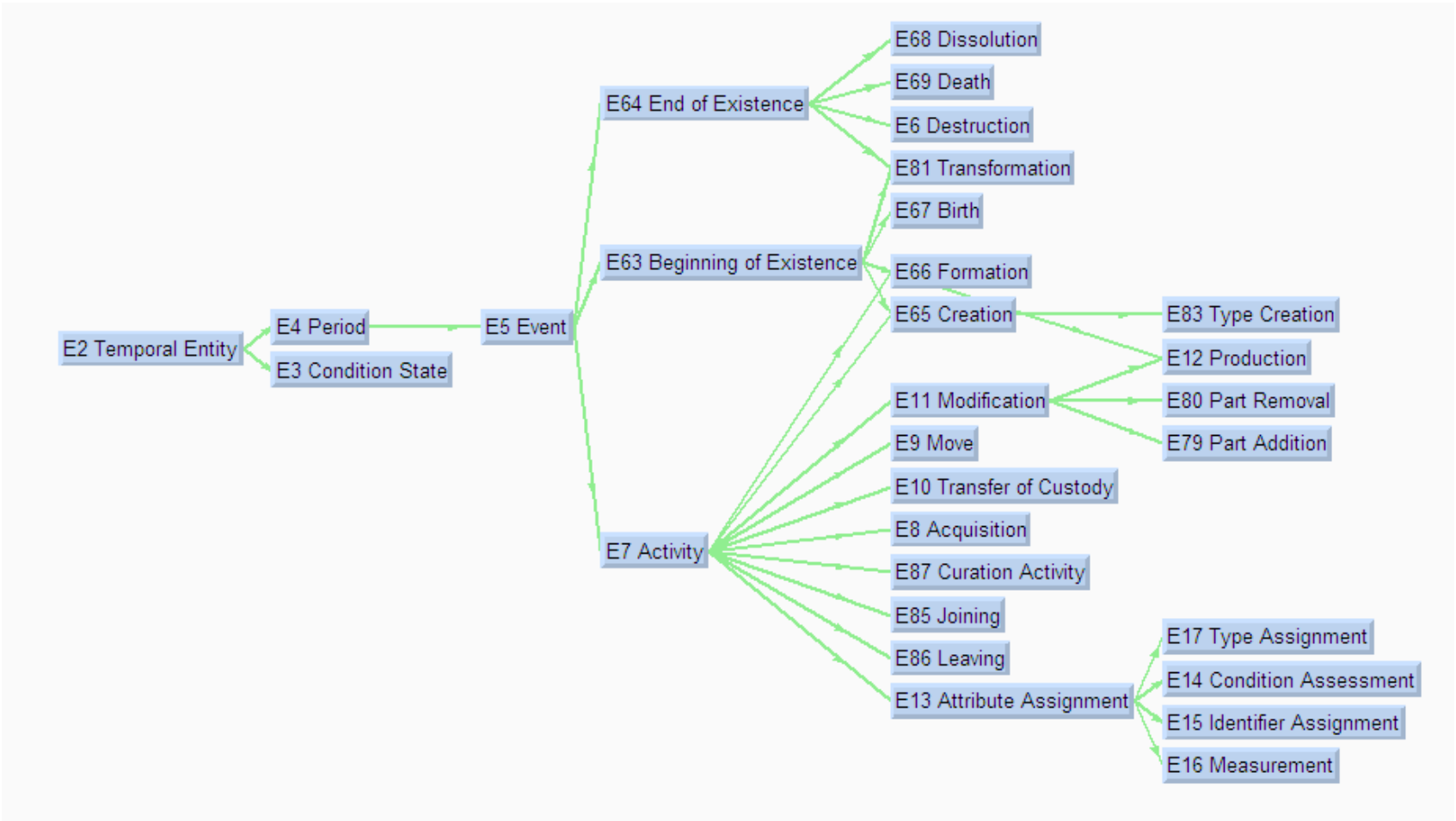
- Dublin Core (general)
- MIDAS heritage standard
- CIDOC Conceptual Reference Model (CIDOC-CRM)

CIDOC-CRM

What?

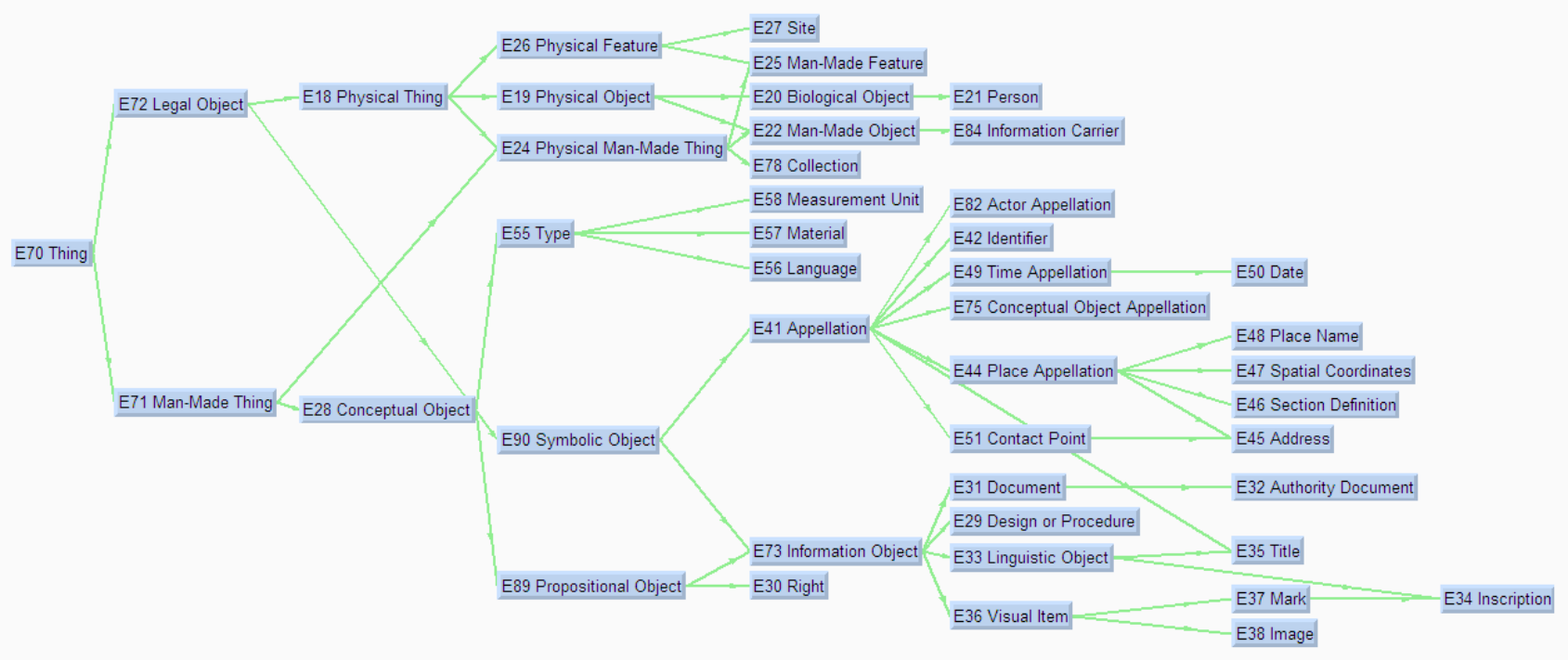
- provides definitions and formal structure for describing concepts and relationships between concepts in cultural heritage
- establishes a formal semantics for the domain description
- can be encoded in XML, RDF(S) etc.

CIDOC-CRM: Temporal Entity Hierarchy



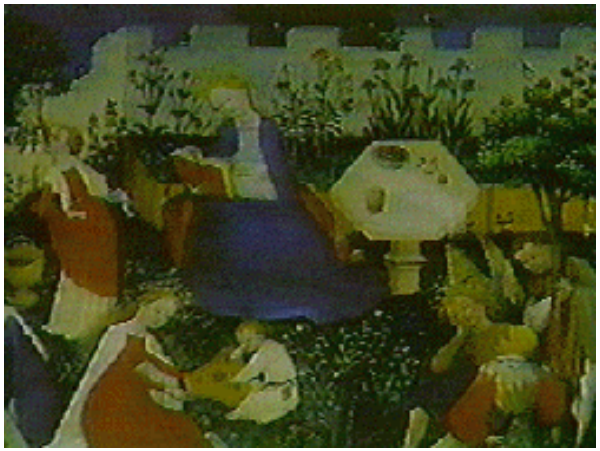
Source: Martin Doerr, Steve Stead, "The CIDOC CRM, a Standard for the Integration of Cultural Information", CRM tutorial at Imperial College, UK, May 22, 2009.

CIDOC-CRM: 'Thing' Hierarchy



Source: Martin Doerr, Steve Stead, "The CIDOC CRM, a Standard for the Integration of Cultural Information", CRM tutorial at Imperial College, UK, May 22, 2009.

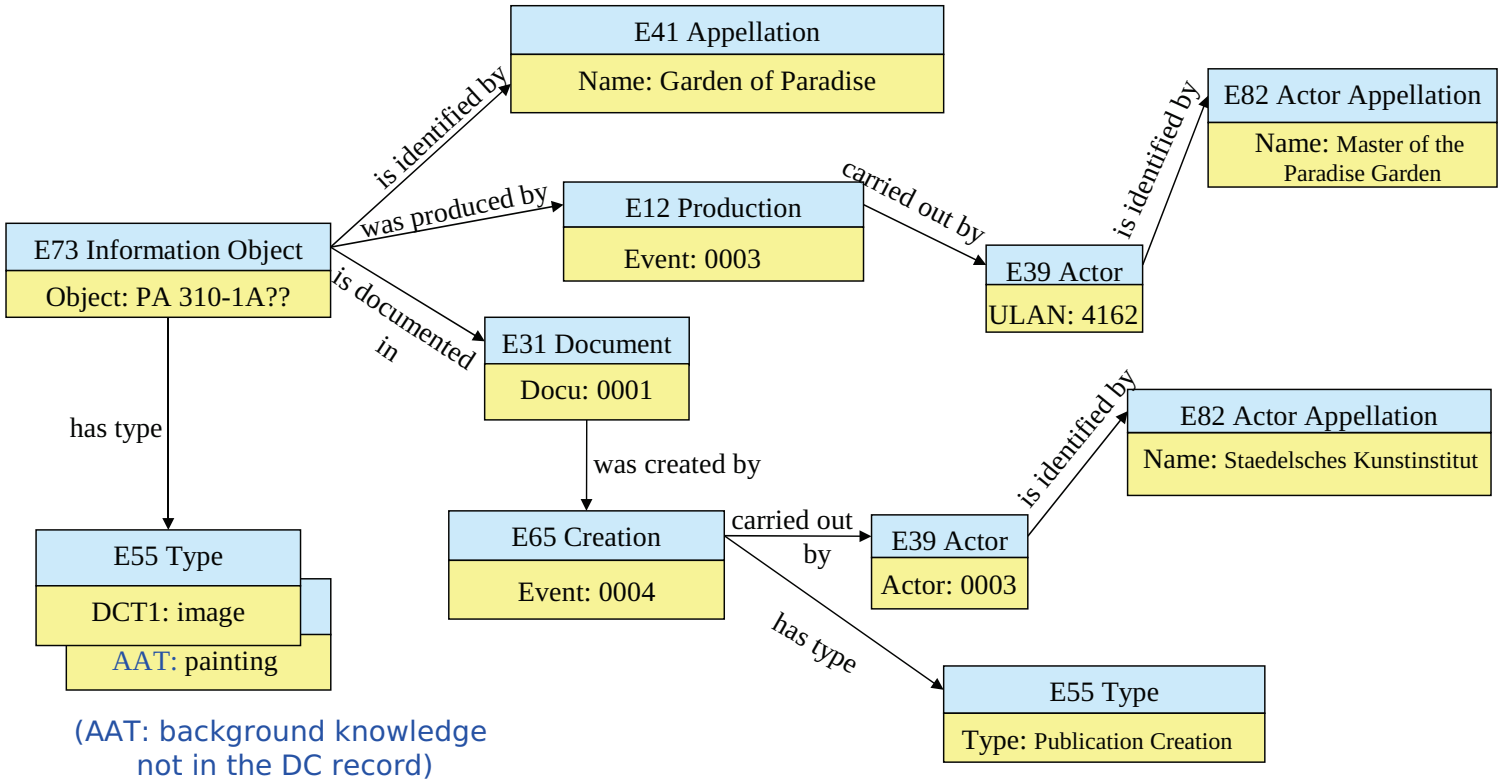
CIDOC-CRM: Example (1)



Type: painting
Title: Garden of Paradise
Creator: Master of the Paradise Garden
Publisher: Staedelsches Kunstinstitut

Source: Martin Doerr, Steve Stead, "The CIDOC CRM, a Standard for the Integration of Cultural Information", CRM tutorial at Imperial College, UK, May 22, 2009.

CIDOC-CRM: Example (2)



Source: Martin Doerr, Steve Stead, "The CIDOC CRM, a Standard for the Integration of Cultural Information", CRM tutorial at Imperial College, UK, May 22, 2009.

Integrating Different Standards

Problem

Standards are good but having many standards doesn't solve the interoperability problem.

Solution

- can map different namespaces, ontologies etc. manually
- RDF provides support for integrating various namespaces
- **but:** automatic mapping would be better
  $\Rightarrow$ still focus of ongoing research (e.g., automatic ontology mapping)